

# **EXHIBIT 5**



OXFORD JOURNALS  
OXFORD UNIVERSITY PRESS

## Society for Political Methodology

---

Iterative Approaches to  $R \times C$  Ecological Inference Problems: Where They Can Go Wrong and One Quick Fix

Author(s): Karen E. Ferree

Source: *Political Analysis*, Vol. 12, No. 2 (Spring 2004), pp. 143-159

Published by: Oxford University Press on behalf of the Society for Political Methodology

Stable URL: <http://www.jstor.org/stable/25791762>

Accessed: 18-02-2018 23:05 UTC

---

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact [support@jstor.org](mailto:support@jstor.org).

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at <http://about.jstor.org/terms>



JSTOR

*Society for Political Methodology, Oxford University Press* are collaborating with JSTOR to digitize, preserve and extend access to *Political Analysis*

# Iterative Approaches to $R \times C$ Ecological Inference Problems: Where They Can Go Wrong and One Quick Fix

Karen E. Ferree

*Department of Political Science, University of California,  
San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0521  
e-mail: keferree@weber.ucsd.edu*

This article argues that a key step in King's iterative approach to  $R \times C$  ecological inference problems—the aggregation of groups into broad conglomerate categories—can introduce problems of aggregation bias and multimodality into data, inducing model violations. As a result, iterative EI estimates can be considerably biased, even when the original data conform to the assumptions of the model. I demonstrate this problem intuitively and through simulations, show the conditions under which it is likely to arise, and illustrate it with the example of Coloured voting during the 1994 elections in South Africa. I then propose an easy fix to the problem, demonstrating the usefulness of the fix both through simulations and in the specific South African context.

## 1 Introduction

Political scientists working on elections in areas of the world where survey data is rare and often flawed received with great interest Gary King's innovative 1997 book on ecological inference problems, hoping it would provide new tools for tackling important empirical questions raised by the spread of democratization to disparate corners of the globe. In fact, a series of recent (or forthcoming) books, articles, and dissertations have used King's technique to do precisely this, applying the method to voting behavior in Zambia (Posner forthcoming), India (Chandra 2004), Poland (Wittenberg and Kopstein 2003), and South Africa (Ferree 2002).

In all of these cases, the data under consideration are complex, involving multiple groups and outcomes. In the language of ecological inference analysis, they form " $R \times C$ " problems: the input data, often population groups, and the output data, typically electoral outcomes, can be arrayed to create a table of multiple rows and columns, where the goal of inference is to estimate the interior cells. This facet of the data would seem to present complications for analysts wishing to apply King's method, because his technique is designed for binary data. However, King suggests that his method can also be used on these more complex problems via an iterative technique wherein the binary model is consecutively applied to subsets of the larger table until all cells of the table can be either estimated or calculated from the values of other cells. In fact, this iterative technique is used to some degree by almost all applied users of King's method.

---

*Author's note:* I thank Gary King, Eric Dickson, Ken Scheve, Neal Beck, and three anonymous reviewers for advice on earlier drafts of this paper. Replication materials are available on the *Political Analysis* Web site.

*Political Analysis*, Vol. 12 No. 2, © Society for Political Methodology 2004; all rights reserved.

This article argues, however, that King's iterative approach to  $R \times C$  ecological inference problems can introduce a set of unanticipated complications that, if unresolved, result in violations of the model's assumptions. The iterative approach involves aggregating groups into broad conglomerate categories like "non-whites" or "non-Christians." Under certain general conditions, lumping together groups in this fashion can introduce aggregation bias and multimodality to the data, even when these problems are not present in the original data.

The goal of this article is to outline the conditions under which the iterative approach is likely to fail, and then to offer ways of circumventing or alleviating the problems that arise when it does. The optimal way for dealing with these problems is to avoid them entirely by estimating all cells of the  $R \times C$  table simultaneously using techniques such as those developed by Rosen et al. (2001). However, until such techniques become easier and more practicable for applied researchers, most people are likely to continue using some variety of the iterative method. For this reason, this article proposes a "quick fix" that can be implemented from within King's framework. The proposed technique performs well in Monte Carlo analysis and appears to significantly improve EI estimates in the real world example explored in this article.

The article is structured as follows: First, I briefly outline King's iterative approach to  $R \times C$  problems. Second, I discuss and explore through Monte Carlo analysis the general conditions under which this causes model failure. Third, I illustrate these points in the context of Coloured voting in South Africa. Fourth, I outline ways of circumventing or resolving the problems that cause model failure, paying special attention to one easy-to-implement remedy.

## 2 The Iterative Approach to $R \times C$ Problems

King's model for ecological inference problems is based on dichotomous data: a binary population variable (e.g., black versus white) as input and a binary behavioral variable (e.g., turnout versus abstain) as output. He assumes that there are no residual categories (for example, people who are neither black nor white, or people who neither vote nor abstain). Of course, as anyone who studies voting knows, the real world tends to present more complex situations. Politically relevant divisions of the population are often not binary but involve many different groups, and interesting political questions tend to involve more than two mutually exclusive behaviors. These situations necessitate a more general  $R \times C$  approach to ecological inference, one that can handle multiple rows and columns of input and output variables. Until practicable  $R \times C$  techniques are developed, however, King suggests that estimates can be produced by iteratively applying his basic model to different subsets of the data.

King uses the example of black, white, and Hispanic turnout to illustrate how the iterative method works. There are three quantities of interest in this case: black turnout for precinct  $i$  ( $\beta_i^b$ ), white turnout for precinct  $i$  ( $\beta_i^w$ ), and Hispanic turnout for precinct  $i$  ( $\beta_i^h$ ). In a straight generalization of the bivariate case, King suggests they can be modeled as a trivariate normal distribution. However, because his model and program are based on the bivariate case, estimating the parameters of this distribution must be done in three steps. First, the white and Hispanic categories are collapsed into a single group and precinct turnout is estimated for blacks ( $\beta_i^b$ ) versus non-blacks ( $\beta_i^{wh}$ ). Second, the white and black categories are collapsed into a single group and precinct turnout is estimated for Hispanics ( $\beta_i^h$ ) versus non-Hispanics ( $\beta_i^{wb}$ ). Third, using the estimates of  $\beta_i^b$  and  $\beta_i^h$ , the final quantity of interest,  $\beta_i^w$ , can be computed deterministically.

### 3 Model Violations

While this process seems unproblematic, difficulties can arise during the first and third steps when the different racial categories (first white and Hispanic and then white and black) are combined into single groups. I discuss how this can induce multimodality and then show how it can also lead to aggregation bias.

#### 3.1 Multimodality

Following King, the precinct-level turnout rate of a residual group can be thought of as a linear combination of the precinct-level turnout rates of its two component groups. For the second case, precinct turnout for the residual white/black group ( $\beta_i^{wb}$ ) is a linear combination of white precinct turnout ( $\beta_i^w$ ) and black precinct turnout ( $\beta_i^b$ ). This can be written:

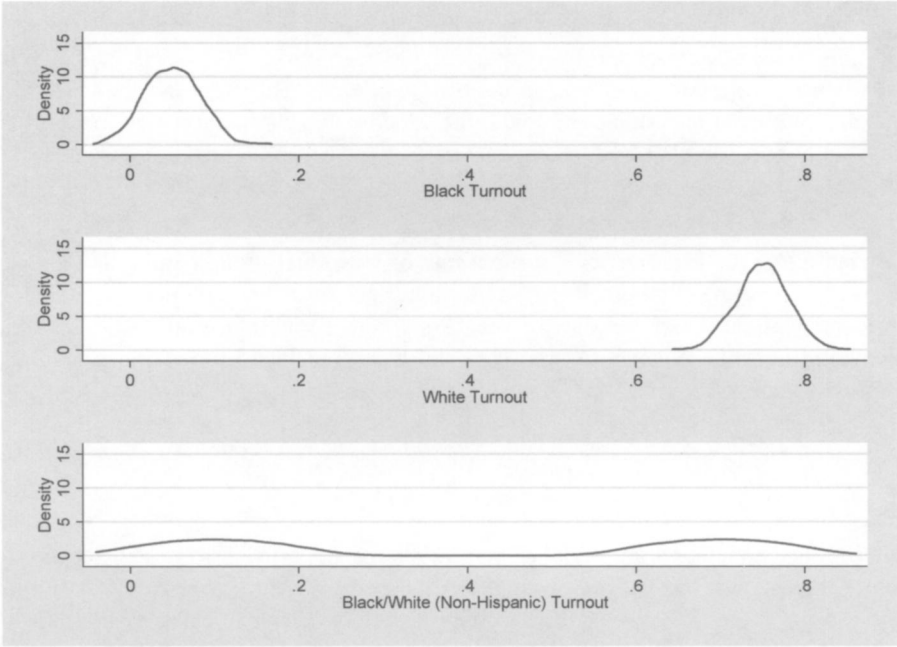
$$\beta_i^{wb} = \left( \frac{X_i^w}{1 - X_i^h} \right) \beta_i^w + \left( \frac{X_i^b}{1 - X_i^h} \right) \beta_i^b.$$

The terms in parentheses are weights: the larger the fraction of whites in the residual category, the greater the weight given  $\beta_i^w$  in the equation and vice versa. To simplify, define  $K_i = X_i^w / (1 - X_i^h)$ . This is the fraction of the non-Hispanic category that is white. The fraction of the non-Hispanic category that is black [ $X_i^b / (1 - X_i^h)$ ] is then  $1 - K_i$ .

The unstated assumption of King's iterative technique is that the precinct level turnout rates for the residual group ( $\beta_i^{wb}$ ) follow a univariate normal distribution. If they do not, then the model's assumption that  $\beta_i^h$  and  $\beta_i^{wb}$  are distributed according to a bivariate normal distribution is problematic.

Assuming that  $\beta_i^w$  and  $\beta_i^b$  are both distributed according to normal distributions, it would seem obvious that  $\beta_i^{wb}$  is also normal, as it appears to be a linear combination of two normally distributed variables. Indeed, in any given precinct,  $\beta_i^{wb}$  is a linear combination of two terms. However, these terms are values of a normally distributed random variable, not normally distributed random variables themselves. Furthermore, the linear combination varies across precincts with  $K_i$ , and there is therefore no single weight term or linear combination. Thus the overall distribution of  $\beta_i^{wb}$  across precincts (which is what is being modeled) is *not* a linear combination of normally distributed variables, and the standard logic about sums of normally distributed variables does not apply.  $\beta_i^{wb}$  will be normally distributed only if  $K_i$  is constant across precincts (thus there is a single weighting or linear combination) or if  $\beta_i^w$  and  $\beta_i^b$  are drawn from the same normal distribution ( $\mu_b = \mu_w$  and  $\sigma_b^2 = \sigma_w^2$ ) so that the weighting becomes irrelevant. The less these conditions hold, the less appropriate the assumption of normality for  $\beta_i^{wb}$ .

Simulations can be used to demonstrate this point. First, 1000 draws of black turnout ( $\beta_i^b$ ) were taken from a normal distribution with mean .05 and variance .001. Then 1000 draws of white turnout ( $\beta_i^w$ ) were taken from a normal distribution with mean .75 and variance .001. These were then combined according to the formula for non-Hispanic turnout ( $\beta_i^{wb}$ ) given above, but the weight term  $K_i$  was allowed to vary across simulations. Half of the simulations were produced using weights that correspond to a precinct in which 5% of the population is black, 65% is white, and 30% is Hispanic; the other half were produced using weights that correspond to a precinct in which 65% of the population is black, 5% is white, and 30% is Hispanic. This experimental situation is therefore one in which the original data do not violate the distributional assumptions of King's model ( $\beta_i^w$  and  $\beta_i^b$  are normal), but do violate the conditions outlined above ( $\beta_i^w$  and  $\beta_i^b$  are not drawn



**Fig. 1** Black, white, and black/white (non-Hispanic) turnout.

from the same distribution and the weights are not constant across precincts). As Fig. 1 shows, the distribution of non-Hispanic turnout ( $\beta_i^{wb}$ ) is clearly not normal.

**3.2** *Aggregation Bias*

Aggregation bias occurs under the same conditions that produce multimodality plus an additional one: correlation between the weight term  $K_i$  and the fraction of the total population that comprises the third population group. This is not a particularly unusual scenario. It will occur, for example, if blacks and Hispanics tend to live in different areas. If this is true, then in precincts with few blacks and lots of Hispanics, the weight on  $\beta_i^b$  will be small and the weight on  $\beta_i^w$  will be large, so the behavior of the conglomerate group  $\beta_i^{wb}$  will heavily reflect white behavior. In contrast, in precincts with lots of blacks but few Hispanics, the weight on  $\beta_i^b$  will be large and the weight on  $\beta_i^w$  will be small, so the behavior of the conglomerate group  $\beta_i^{wb}$  will heavily reflect black behavior. Thus  $\beta_i^{wb}$  will vary with the size of the Hispanic population, producing aggregation bias—even if this was not present in the original data.

In sum, problems arise when some or all of the following conditions fail to hold:

- 1.  $K_i = K_j$  for all  $i$  and  $j$ ;
- 2.  $\mu_b = \mu_w$  and  $\sigma_b^2 = \sigma_w^2$ ;
- 3.  $cor(K_i, X_i) = 0$ .

Condition 1 requires that the weight term  $K_i$  be constant across precincts. We can think of this as a precinct *homogeneity* assumption: the precincts must be homogeneous mixtures of groups. Condition 2 requires that the groups being combined have identical

behavior. We can think of this as an *absence of polarization* assumption. Condition 3 requires that there be no correlation between the weight term  $K_i$  and the size of the population groups whose behavior is being estimated (in the running example, Hispanics and non-Hispanics). We can think of this as an *absence of correlation* assumption. If conditions 1 and 2 both fail, then collapsing groups into general categories produces multimodality. If conditions 1, 2, and 3 fail, then collapsing groups produces multimodality and aggregation bias.

More intuitively, the assumption underlying King's model is that the behavior of a group in precinct  $i$  has "something in common" with the behavior of a group in precinct  $j$ . If the group in question is a composite of two groups that behave quite differently, then its behavior will depend on the amount of each group in it. Where this fraction varies considerably across precincts, the assumption that all precincts have something in common becomes difficult to support.

In the running example, the composite group "non-Hispanics" consists of blacks and whites. If most blacks vote while whites hardly vote at all, *and* some precincts are mostly white while others are mostly black, then some precincts will have a high non-Hispanic turnout rate and others will have a low one. They will not all have something in common. Furthermore, if it is also true that blacks and Hispanics tend to live in the same precincts, then precincts with many Hispanics will also be precincts in which black behavior weights strongly in the behavior of non-Hispanics, causing turnout to be higher. Thus the fraction of the total population that is Hispanic will correlate with the behavior of non-Hispanics, thereby inducing aggregation bias.

Note that problems with multimodality and aggregation bias can arise for collapsed groups *even if they are not present in the original data*. That is,  $\beta_i^w$ ,  $\beta_i^b$ , and  $\beta_i^h$  may follow a trivariate normal distribution. Furthermore, all three may be uncorrelated with  $X_i^w$ ,  $X_i^b$ , and  $X_i^h$ . However, this does not imply that  $\beta_i^h$  and  $\beta_i^{wb}$  follow a bivariate normal distribution, nor does it imply that  $X_i^h$  is uncorrelated with  $\beta_i^{wb}$ . Thus diagnosing situations in which aggregating groups generate new problems requires thinking about the data in new ways.

On a more positive note, two of the three conditions listed above (the first and third) are easily diagnosable from aggregate data. Furthermore, researchers typically have some intuition about the degree of behavioral convergence across groups (the second condition), even if they do not have point estimates for the behavior. For these reasons, the problems outlined in this paper are relatively easy to identify with basic contextual knowledge—unlike the more difficult problem of aggregation bias in the original data.

#### 4 Monte Carlo Simulations

Monte Carlo simulations permit a more thorough exploration of the problems created for EI by the failure of one or more of the conditions outlined in Section 2. In the following section, I explore how different degrees and combinations of heterogeneity, polarization, and correlation affect the quality of EI estimates. In all scenarios, Hispanic, white, and black turnout levels were drawn from normal distributions. Thus the underlying data fit the assumptions of EI. Each scenario was run 100 times and had a sample size of 200 precincts.

First, I explored EI's performance when there is high heterogeneity, low polarization, and varying levels of correlation. Thus  $K_i$  and  $X_i$  were drawn from a bivariate normal distribution. The variance of  $K_i$  was set high [ $K_i \sim N(.2, .0625)$ ], corresponding to a situation of significant heterogeneity. Across different sets of simulations, the correlation

Table 1 Monte Carlo simulations

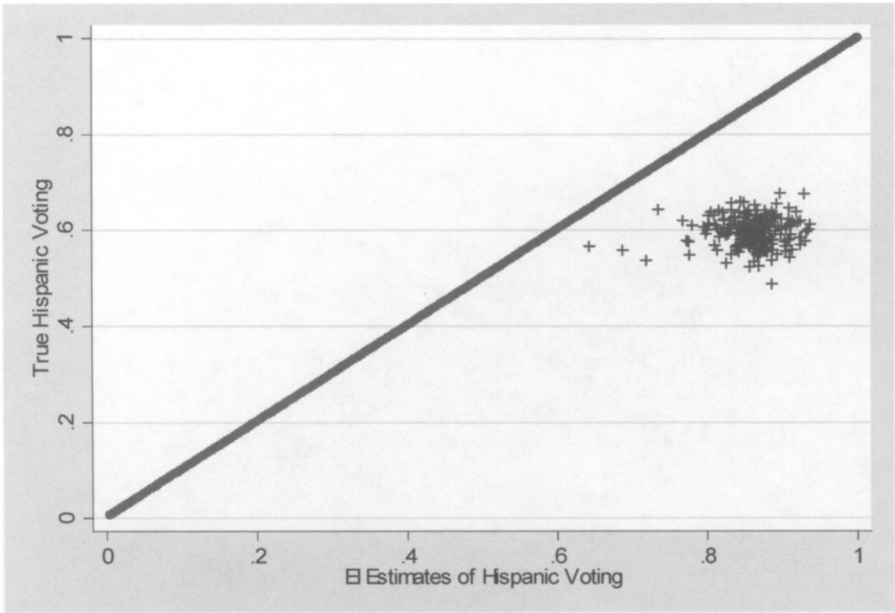
	"Hispanic" voting			"Non-Hispanic" voting		
	EI	TRUTH	DIFF	EI	TRUTH	DIFF
High heterogeneity, low polarization						
Rho = .9	.6030	.6003	.0050	.4976	.4999	.0038
Low heterogeneity, high polarization						
Rho = .9	.7031	.6001	.1052	.3448	.4270	.0836
High heterogeneity, high polarization						
Rho = .1	.6137	.6003	.0168	.4135	.4260	.0154
.3	.6433	.6001	.0433	.3873	.4263	.0390
.6	.7115	.6001	.1115	.3314	.4274	.0960
.9	.8366	.6003	.2364	.2345	.4273	.1928

Note. Based on 100 simulations of 200 precincts. Estimates are of aggregate– (not precinct–) level behavior. Rho: correlation of  $X_{ic}$  (Hispanic population) and  $K_i$  (fraction of non-Hispanic population that is white). High variance:  $K_i \sim N(.2, .0625)$ ; low variance:  $K_i \sim N(.2, .0100)$ . High polarization: mean of black turnout = .05, mean of white turnout = .80. Low polarization: mean of black turnout = .45, mean of white turnout = .55.

between  $K_i$  and  $X_i$  was set at various levels ranging from very low (.1) to very high (.9). The polarization between white and black turnout was set low [ $\beta_i^w \sim N(.55,.001)$  and  $\beta_i^b \sim N(.45,.001)$ ]. Under these conditions (violation of conditions 1 and 3, satisfaction of condition 2), EI performs quite well. Table 1, first line, shows the results for the highest level of correlation between  $K_i$  and  $X_i$ . This is where we would expect the biggest problems because condition 3 is most severely violated. However, EI zoned in quite precisely on the truth for both Hispanic and non-Hispanic voting. Thus, when polarization is low, heterogeneity of weights and correlation between weights and population sizes do not create problems. This makes perfect sense: different mixtures of like things will not vary much, so multimodality and aggregation bias generated by combining categories will be extremely mild.

Second, I explored EI’s performance when there is low heterogeneity, high polarization, and varying levels of correlation. The variance of  $K_i$  was set low [ $K_i \sim N(.2,.0100)$ ], corresponding to a situation of relative homogeneity of mixtures across districts. Across different sets of simulations, the correlation between  $K_i$  and  $X_i$  was set at various levels ranging from very low (.1) to very high (.9). The polarization between white and black turnout was set high [ $\beta_i^w \sim N(.80,.001)$  and  $\beta_i^b \sim N(.05,.001)$ ]. As the second line of Table 1 shows, under these conditions (violation of conditions 2 and 3, satisfaction of condition 1), EI begins to have trouble at high levels of correlation between  $K_i$  and  $X_i$ : it misses the truth for Hispanic voting by about 10 percentage points and misses the truth for non-Hispanic voting by about 8 percentage points. At lower levels of correlation (not shown), the gap between the truth and EI estimates is much milder. Thus, when the groups being aggregated into the combined category have sharply different behavior, even low levels of heterogeneity (when combined with correlation) can create problems.

Third, I explored the worst–case scenario: heterogeneity, polarization, and varying levels of correlation. The variance of  $K_i$  was set high [ $K_i \sim N(.2, .0625)$ ], corresponding to a situation of significant heterogeneity. Once again, the correlation between  $K_i$  and  $X_i$  was set at various levels ranging from very low (.1) to very high (.9). The polarization between white and black turnout was set high [ $\beta_i^w \sim N(.80,.001)$  and  $\beta_i^b \sim N(.05,.001)$ ]. As the last

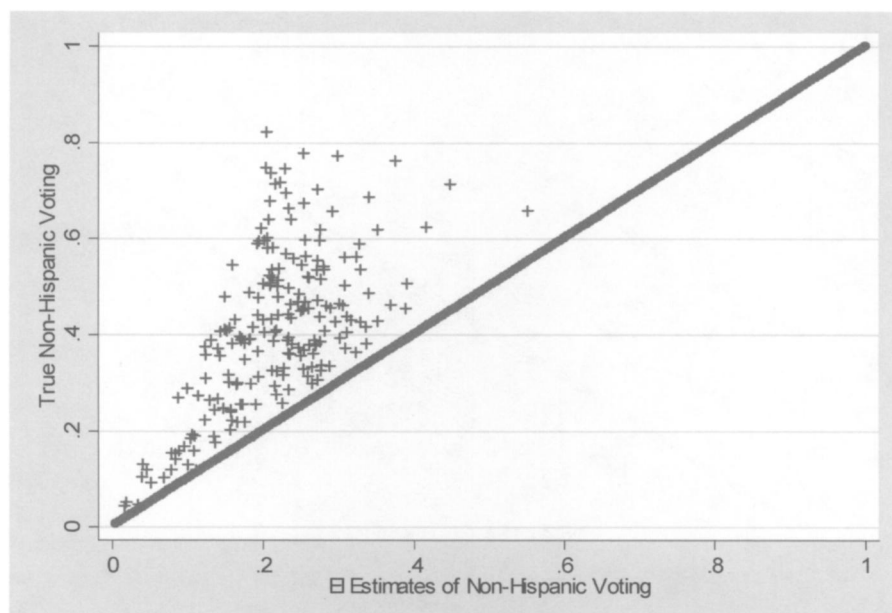


**Fig. 2** EI precinct-level estimates of Hispanic voting versus true precinct-level values, based on a worst-case scenario: high heterogeneity, high polarization, high correlation.

four rows of Table 1 show, EI’s estimates diverge from the truth in progressively larger degrees as the correlation between  $K_i$  and  $X_i$  increases. At mild levels of correlation (.3), EI is missing the truth for Hispanic and non-Hispanic voting by about 4 percentage points. This increases to 10 to 11 percentage points when correlation is moderate (.6). When correlation is high (.9), EI’s errors are severe: around 24 percentage points for Hispanic aggregate turnout and 20 percentage points for non-Hispanic aggregate turnout. It bears repeating that these errors occur *even though the underlying behavioral data conform to the model*; they are due entirely to the aggregation of unlike groups in the combined non-Hispanic category.

The results in Table 1 show only aggregate quantities of interest. However, we often care more about precinct-level quantities of interest. Therefore it is useful to examine EI’s performance at the precinct level for the worst-case scenario presented above: heterogeneity, polarization, and high correlation (the last line in Table 1). Figure 2 plots EI precinct-level estimates of Hispanic turnout against the truth. If EI estimates matched the truth, then all of the points should line up along the 45° line. However, the points form a cluster well off the 45° line, visually demonstrating EI’s precinct-level errors. Figure 3 shows a similar plot of EI’s precinct-level estimates of non-Hispanic turnout against the truth. Once again, the extent of EI’s errors is clear: the points spray up and out from the 45° line.

Thus, at the precinct level, as at the aggregate level, EI’s estimates diverge significantly from the truth under the conditions identified in Section 3: heterogeneity of mixtures across precincts, polarization in the conglomerate category, and correlation between  $K_i$  and  $X_i$ . Simulations therefore demonstrate conclusively that the iterative technique for estimating  $R \times C$  problems can induce violations in EI’s assumptions, even when these assumptions are not violated in the underlying data.



**Fig. 3** EI precinct-level estimates of non-Hispanic voting versus true precinct-level values, based on a worst-case scenario: high heterogeneity, high polarization, high correlation.

## 5 Application: Coloured Voting During the 1994 South African National Election

How do these issues play out in real data? An example drawn from recent South African electoral history provides insights into this. The voting behavior of Coloured people during the 1994 South African national election has been the subject of much writing and speculation.<sup>1</sup> Seven parties won seats in the 1994 elections: the African National Congress (ANC), the National Party (NP), the Inkatha Freedom Party (IFP), the Democratic Party (DP), the Pan African Congress (PAC), the Freedom Front (FF), and the African Christian Democratic Party (ACDP). All available evidence indicates that African and white voting was extremely polarized during the 1994 election (and all elections since). Africans supported one set of parties—ANC, IFP, and PAC—while whites supported a different and nonoverlapping set of parties—NP, DP, and FF.<sup>2</sup> There was virtually no cross-racial voting, leading many to describe the election in Horwitzian terms as an “ethnic census” (Mattes 1995; Johnson and Schlemmer 1996). Unlike the white and African vote, however, the Coloured vote was “up for grabs,” producing intense competition on the part of several political parties, including the ANC and the NP.

Consequently there have been several attempts to estimate Coloured voting, most of which suggest that around 60–70% of voting Coloureds supported the NP (Mattes et al. 1996; Seekings 1996). Because exit polls were prohibited in the 1994 elections, these

<sup>1</sup>Coloureds, one of South Africa’s four major population groups, make up about 8% of the South African population. They are heavily concentrated in the Western and Northern Cape provinces and some regions of the Eastern Cape. Analyses of Coloured voting in the 1994 election include Mattes (1995); Mattes et al. (1996); and Seekings (1996).

<sup>2</sup>The only exception to this is in KwaZulu-Natal (one of South Africa’s nine provinces), where a small fraction of whites voted for the IFP. For data reasons (suspected high levels of fraud), KwaZulu-Natal was dropped from this analysis. This is not problematic from the perspective of estimating Coloured voting because virtually no Coloureds live in this province.

estimates are based on data from national and provincial surveys conducted at least two months prior to the election when many voters were undecided about which party they planned to support. Ecological inference techniques, by providing estimates of Coloured behavior at the disaggregated level of magisterial districts (the lowest level for which both electoral and census data were available in 1994) and by using the actual output of the election itself, may therefore provide an interesting new window on an old question.

Estimating Coloured behavior using King's technique requires combining Africans and whites into a single aggregate "non-Coloured" category.<sup>3</sup> King's method can then be used to estimate turnout for Coloureds and non-Coloureds; these estimates can in turn be used to estimate Coloured and non-Coloured support for the National Party.<sup>4</sup>

A priori, how likely is it that the creation of the non-Coloured category will cause problems for these data? The answer is different for turnout versus support for the National Party. Condition 1 does not hold for South Africa: the composition of non-Coloureds varies significantly across magisterial districts. While one-quarter of the districts have non-Coloured populations that are 100% African, around 12% of the districts have non-Coloured populations that are least 60% white. Around one-third of the districts occupy a middle range, with non-Coloureds fairly balanced between Africans and whites.

Condition 3 also does not hold for South Africa: Coloureds and Africans tend to live in different parts of the country, ensuring that the fraction of the non-Coloured population that is white correlates with the fraction of the total population that is Coloured. More specifically, the correlation coefficient is .83.

Condition 2, on the other hand, seems likely to hold for turnout. Prior studies suggest that all population groups turned out to vote at very high rates during the historic 1994 "liberation" elections, and there is little evidence suggesting that turnout rates varied significantly across groups (Reynolds 1994). The opposite case holds for support for the National Party. As discussed above, whites and Africans diverge considerably in their patterns of party support. Survey evidence prior to the election suggested that around 44% of whites would vote for the NP (with 29% undecided). In contrast, only 1% of Africans indicated they would support the party. Postelection estimates suggest even more skewed support rates, with around two-thirds of whites supporting the NP versus less than 2% of Africans (Mattes et al. 1996; Seekings 1996).

In sum, there are strong reasons to believe that the non-Coloured category will pose few problems for estimating turnout but major problems (multimodality and aggregation bias) for estimating support for the National Party.

Following King's notation, Coloured turnout is  $\beta_i^c$ ; non-Coloured turnout is  $\beta_i^{aw}$ ; Coloured support for the NP is  $\lambda_i^c$ ; and non-Coloured support for the NP is  $\lambda_i^{aw}$ . Table 2 shows the results of these estimations.

The first-round estimates show high turnout rates for both Coloureds and non-Coloureds and are consistent with those given in other studies.<sup>5</sup> The estimates of Coloured support for the NP, however, are 14 to 20 percentage points *higher* than those suggested

<sup>3</sup>South Africa also has a fourth major population (i.e., racial) group—Indians/Asians. They form only around 3% of the entire population and are not considered separately in this analysis. Outside of KwaZulu-Natal, which is omitted from the analysis due to data quality issues, the fraction of non-Coloureds that is Indian is negligible (less than 1%) and therefore has little impact on the results.

<sup>4</sup>Alternatively, one could estimate support for the ANC.

<sup>5</sup>The Independent Electoral Commission's publication *The South African Elections of April 1994* gives an overall turnout rate of 86.5%. The estimates above are somewhat low as a number of districts had to be cut from analysis because turnout exceeded 100%—a reflection of poor census data in some areas of the country. Because most of these areas are heavily African, they have little impact on the estimation of Coloured behavior. They do, however, depress the estimates of non-Coloured turnout slightly.

**Table 2** Initial estimates of Coloured and non-Coloured turnout and support for the National Party in South Africa’s 1994 election (percentages)

<i>Parameter</i>	<i>Estimate</i>	<i>Standard error</i>
Coloured turnout ( $\beta_i^c$ )	.86	.02
Non-Coloured turnout ( $\beta_i^{aw}$ )	.82	.00
Coloured votes for NP ( $\lambda_i^c$ )	.84	.02
Non-Coloured votes for NP ( $\lambda_i^{aw}$ )	.14	.00

*Note.*  $N = 216$ . Districts where Coloureds are less than 1% of the population and the National Party won less than 5% of the vote are selected out of the estimation stage. Based only on Eastern Cape, Northern Cape, and Western Cape.

by survey research: a substantial discrepancy. Due to this discrepancy (as well as initial suspicions that the non-Coloured category would pose more problems for estimating party choice versus turnout), the remainder of this section focuses on this second set of results.

Standard EI diagnostics show little evidence of the model not fitting the data. A scatterplot of the fraction of Coloureds in the total population versus support for the NP shows relatively narrow 80% confidence intervals, within which at least 80% of the points lie. King’s “tomography” plots with maximum likelihood contours also suggest good fit. Virtually all of the tomography lines overlap in one small area of the graph, suggesting a single mode for the distribution. In short, standard EI diagnostics indicate good fit.

Table 3, which selects out 10 districts scattered around the Eastern, Western, and Northern Cape provinces (areas with the densest settlements of Coloureds), reveals a different story. These districts vary considerably: Some have small Coloured populations (Bedford), some large (Uniondale). In some, the non-Coloured population is mostly African (Bedford and Nouport); in some it is mostly white (Calizdorp, Sutherland, Bellville, Hopefield). Estimates for Coloured voting ( $\lambda_i^c$ ) also vary considerably, from .60 in Graaff-Reinet to .97 in Hopefield.

The most revealing aspect of this table is the estimates for non-Coloured voting ( $\lambda_i^{aw}$ ). These estimates vary over the districts (from .06 in Graaff Reinet to .56 in Bellville), *but not as much as they should*.

As mentioned earlier, whites voted exclusively for three parties: the National Party, the Democratic Party, and the Freedom Front. As shown in the table, the combined total vote for the DP and the FF rarely exceeded 5%. Therefore, districts in which the non-Coloured population is virtually all white should have very high rates of support for the National Party. Yet the highest support rate in the table is 56% in Belleville. In a district in which whites form 90% of the non-Coloured population and the DP and FF combined won only 7% of the vote, this is unrealistic. Even more unrealistic are the estimates for Calizdorp (where whites form 96% of non-Coloureds) and Sutherland (where they form 93%). EI estimates that only 19% of non-Coloureds in Calizdorp and 38% in Sutherland voted for the NP. Even the upper limits of the confidence intervals for these districts are unrealistic. Where are the rest of the white votes going, if not for the NP, the DP, or the FF? There is nothing to suggest that these areas have some peculiar and exceedingly rare clump of liberal ANC-supporting whites. In fact, if anything, the opposite should be true—these districts are mostly rural and the culture is decidedly traditional and conservative, not fertile hunting ground for African parties. Looking at the table in this light, virtually all of the estimates for non-Coloured support for the NP seem suspiciously low.

Closer consideration of Calizdorp reveals the extent of the problem. Approximately 2900 Coloureds, 800 whites, and 35 Africans cast votes during the 1994 elections in this

**Table 3** Various district-level estimates of support for the National Party in the 1994 South African election

District	White		Votes won by NP	Votes won by DP/FF	Coloured NP support	Non-Coloured NP support
	non-Coloured voters	Coloured voters				
Bedford (EC)	.10	.16	.22	.03	.82 (.64,.94)	.07 (.03,.11)
Noupoort (NC)	.30	.37	.43	.04	.86 (.75,.96)	.17 (.10,.23)
Aberdeen (EC)	.53	.69	.56	.07	.77 (.71,.82)	.09 (.02,.17)
Graaff-Reinet (EC)	.44	.61	.40	.03	.60 (.56,.64)	.06 (.00,.13)
Victoria West (NC)	.59	.73	.73	.05	.94 (.90,.99)	.25 (.13,.35)
Hopefield (WC)	.90	.60	.75	.07	.97 (.93,.99)	.42 (.37,.49)
Belleville (WC)	.90	.40	.73	.07	.96 (.89,1.00)	.56 (.50,.62)
Calizdorp (WC)	.96	.78	.78	.06	.92 (.87,.97)	.19 (.07,.32)
Uniondale (WC)	.75	.82	.63	.05	.72 (.67,.78)	.13 (.02,.25)
Sutherland (NC)	.93	.71	.81	.09	.98 (.95,1.00)	.38 (.25,.52)

Note. NC = Northern Cape, EC = Eastern Cape, WC = Western Cape. Estimates of  $\lambda_i^c$  and  $\lambda_i^{aw}$  show 80% confidence intervals.

remote Western Cape district. The NP won a total of 2909 votes, the DP 35, and the FF 174. EI estimates that 19% of the 835 non-Coloureds voted for the NP—around 160 votes. Assuming that no Africans voted for the NP, this implies that something like 640 whites did *not* vote for the NP. Subtracting votes for the FF and the DP, EI is estimating that over half the whites in this district did not vote for one of the three white parties. Even taking into account the range implied by the confidence intervals, these are extremely unlikely results for a rural area known for its conservatism. Calizdorp, though the most obvious problem in Table 3, is by no means an outlier.

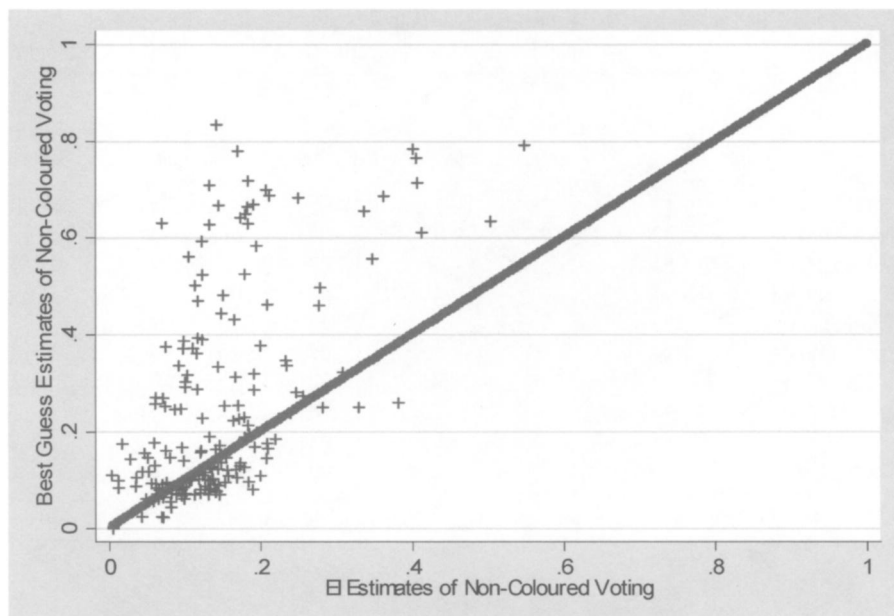
Figure 4 extends this analysis to all of the districts by plotting a “best guess” for non-Coloured support for the NP by EI’s estimate. The best guess is arrived at through a logic similar to that outlined above: It is assumed that only whites voted for the DP and FF.<sup>6</sup> These votes are therefore subtracted from the total number of white voters. All remaining white voters are assumed to have voted for the NP; African voters are assumed not to have voted for the NP at all. Therefore white voters not voting for the DP or the FF are divided by the total number of non-Coloured voters to give an estimate of the fraction of non-Coloureds supporting the NP.

The EI estimates deviate significantly from common sense: very few of the confidence intervals intersect the 45° line. Furthermore, EI appears to be systematically under-estimating non-Coloured support for the National Party. In sum, contextual evidence suggests that iterative EI is not doing a very good job of estimating Coloured and non-Coloured support for the National Party.

6 Avoiding and/or Fixing the Problem

There are several routes around the problems generated by the iterative  $R \times C$  method. The best method is to estimate the cells of the  $R \times C$  table simultaneously rather than iteratively. Rosen et al. (2001) propose a multinomial-Dirichlet ecological inference model

<sup>6</sup>It is unlikely that *any* nonwhites voted for the Freedom Front due to their emphasis on white (Afrikaner) rights. Some nonwhites probably voted for the Democratic Party, but the numbers were small enough to have a low impact on the analysis.



**Fig. 4** EI estimates of non-Coloured voting versus best guess estimates (South Africa data).

that does this, and two methods for estimating it: a Bayesian approach that uses Markov chain Monte Carlo (MCMC) methods and a frequentist approach based on a first moments estimator. They suggest that these methods are complementary: the full Bayesian approach is richer but more difficult to implement, while the first moments approach is easier to implement but does not provide estimates of all quantities of interest and is less efficient. In a later piece, King et al. (2003) recommend dealing with this trade-off by using the first moments approach to get point estimates and the MCMC approach to generate standard errors.

While these methods are promising and avoid the problems highlighted in this paper by estimating the entire  $R \times C$  table at once, they are not currently practicable for the majority of researchers studying elections. As a result, most analyses to date continue to rely on the iterative technique, warts and all. For these reasons, until the Rosen et al. approach is easier to implement, “quick fixes” that can be implemented within the framework of King’s method are likely to be of greater value. Below, I outline two such fixes.

### 6.1 Estimation Order in Iterative Techniques

Depending on the particular research situation, changing the order in which the cells of the  $R \times C$  table are estimated may eliminate or reduce the severity of aggregation bias and/or multimodality caused by collapsing rows or columns of  $R \times C$  tables. In the example of Coloured voting in South Africa, estimating African vs. non-African votes and then breaking non-African votes into Coloured and white votes could improve on the results presented earlier (where the data were first broken down as Coloured vs. non-Coloured). Survey evidence suggests that Coloureds and whites supported the National Party at similar rates in the 1994 election, somewhere in the 60–70% range. Therefore, aggregating Coloureds and whites into a “non-African” group may present fewer problems than aggregating whites and Africans into a “non-Coloured” group. The behavior of

Table 4 Monte Carlo simulations for raw and fixed data

	"Hispanic" voting			"Non-Hispanic" voting		
	EI	TRUTH	DIFF	EI	TRUTH	DIFF
High polarization						
Raw	.8366	.6003	.2364	.2345	.4273	.1928
Fixed <sup>a</sup>	.6101	.6003	.0116	.4192	.4273	.0154
Low polarization						
Raw	.6030	.6003	.0050	.4976	.4999	.0038
Fixed <sup>a</sup>	.6021	.6003	.0077	.4987	.5000	.0058

Note. Based on 100 simulations of 200 precincts. All high variance, high rho (see Table 1). Estimates are of aggregate (not precinct) level behavior. High polarization: mean of black turnout = .05, mean of white turnout = .80. Low polarization: mean of black turnout = .45, mean of white turnout = .55.

<sup>a</sup>Fixed using  $K_i$  (size of non-Hispanic population that is white) covariate, with a prior of (00, .01).

non-Coloureds depends heavily on the ratio of the groups composing it; this is less true for non-Africans.

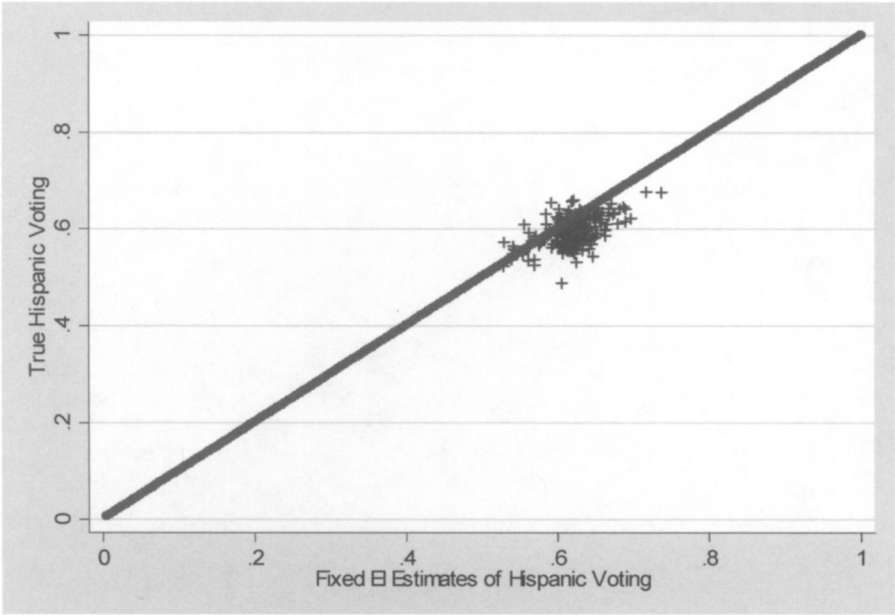
6.2 Using a Covariate to Control for Aggregation Bias and/or Multimodality

In situations in which the original  $R \times C$  table is relatively small, it is possible to control for the effects of collapsing rows or columns by including a covariate that measures the weight of each group in the combined category ( $K_i$ ). The justification for this technique is based on King’s (1997) recommendations for dealing with aggregation bias and multimodality. He suggests that in both instances a covariate can improve EI’s performance. In the case of multimodality, he recommends a covariate that captures the factor or factors that cause districts to differ (pp. 187–188). With aggregation bias, he recommends using a covariate that captures the correlation between the precinct-level quantities of interest and the size of the population groups (pp. 174–179). When multimodality and aggregation bias are a function of the same underlying process (as they are in the examples pursued in this article), then one covariate can fix both problems. In this case, a covariate that measures the weight of each group in the combined category ( $K_i$ ) does precisely this. Thus the justification for this technique comes directly from King.<sup>7</sup>

To evaluate the effects of using a covariate, I looked at two scenarios. The first is the worst-case scenario detailed in Section 3: high heterogeneity, high polarization, and high correlation. The second scenario replicates the first scenario, only polarization is set low. These data were first run using straight EI (the “raw” data) and were then run using the covariate (the “fixed” data). Comparing the output of these four different sets of EI runs (each based on 100 simulations of 200 precincts, as before) allows us to assess first, if the covariate improves the functioning of EI when data are polarized, and second, if the covariate diminishes the functioning of EI when the data are not polarized.

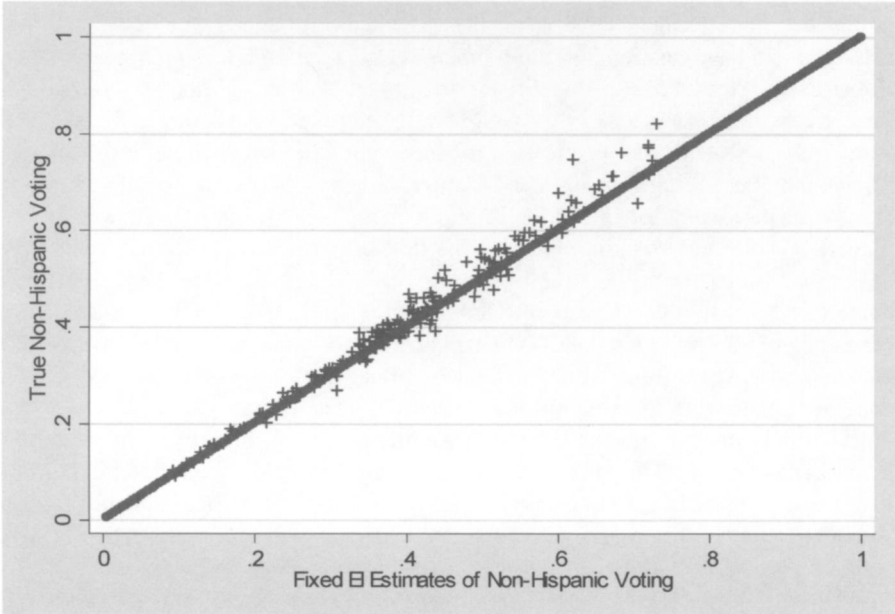
Table 4 presents the results of these simulations. The first two rows, the high polarization case, suggest that the covariate significantly improves the functioning of EI. Without the covariate, EI misses by around 23 percentage points for Hispanic voting and 19 percentage points for non-Hispanic voting. With the covariate, EI misses by one to two

<sup>7</sup>It is worth noting that the problems generated by the iterative  $R \times C$  method (multimodality and aggregation bias) are not themselves novel problems; King (1997) discusses them in great detail. What is novel here is the way in which these problems are generated.



**Fig. 5** Fixed precinct-level EI estimates of Hispanic turnout versus true precinct-level values.

percentage points: a large improvement. This improvement is also obvious when we examine graphs of EI’s fixed precinct-level estimate plotted against the true precinct-level values. Figure 5 shows that Hispanic precinct-level estimates cluster over the 45° line—a clear improvement over Fig. 2. Similarly, Fig. 6 shows that the spray of non-Hispanic precinct-level estimates follows the 45° line—a clear improvement over Fig. 3. Thus, at



**Fig. 6** Fixed precinct-level EI estimates of non-Hispanic turnout versus true precinct-level values.

**Table 5** Estimates of coloured and non-Coloured support for the National Party in South Africa’s 1994 election, with covariate

<i>Parameter</i>	<i>Estimate</i>	<i>Standard error</i>
Coloured votes for NP ( $\lambda_i^c$ )	.64	.01
Non-Coloured votes for NP ( $\lambda_i^{aw}$ )	.16	.00

*Note.*  $N = 216$ . Districts where Coloureds are less than 1% of the population and the National Party won less than 5% of the vote are selected out of the estimation stage. Based only on Eastern Cape, Northern Cape, and Western Cape.

the aggregate and precinct levels, the covariate seems to cure the problems introduced by the iterative method.

Furthermore, as the second two rows of Table 2 show, including a covariate when the underlying data are *not* polarized introduces no new problems: for both the raw and fixed data, EI is within a percentage point of the truth. In situations in which the data clearly violate conditions 1 and 3, but polarization is unknown, researchers can therefore use a covariate without worrying that doing so might create a new foul.

In the case of Coloured voting in South Africa, including the size of the African voter population as a covariate also considerably improves results. Table 5 shows the aggregate results of these estimations. Including the covariate causes the estimate of Coloured support for the NP to drop by 20 percentage points. The new estimate is more in line with survey results. The estimate for non-Coloured support for the NP increases a little—by two percentage points. Table 6 shows district-level results, which show even more dramatic changes for some areas.

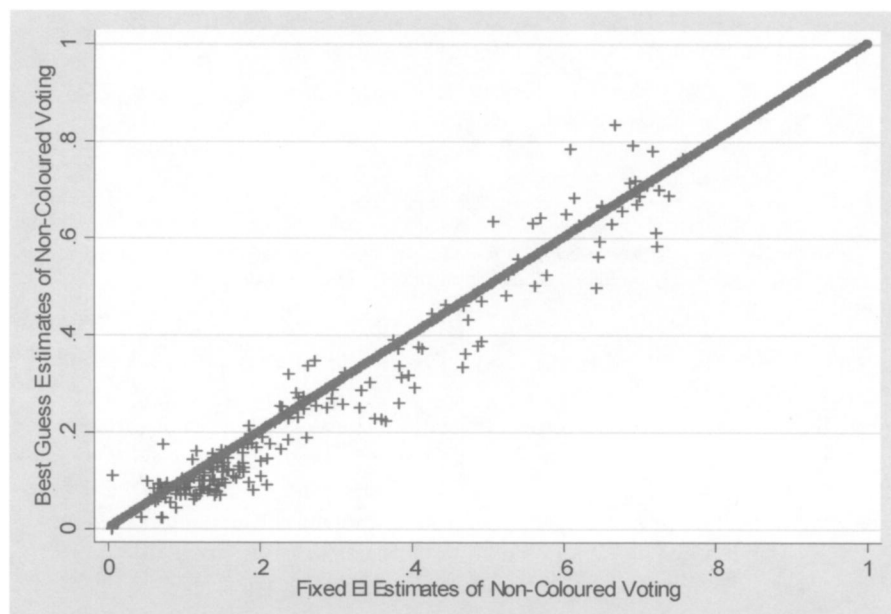
In Calizdorp, our previous example, the non-Coloured estimate increases by over 50 percentage points, much more in line with common sense for that district. Several other districts increase by around 30 percentage points. Coloured estimates, though not off by as much, all drop by 10 to 20 percentage points. These are not minor changes.

Figure 7 plots out the “best guess” against the new estimates for all precincts. The scatterpoints line up reasonably well with the 45° line, suggesting that the fixed estimates

**Table 6** Various district-level estimates of support for the National Party in the 1994 South African election, with covariate

<i>District</i>	<i>Coloured NP support</i>	<i>Change from old estimates</i>	<i>Non-Coloured NP support</i>	<i>Change from old estimates</i>
Bedford (EC)	.63 (.57,.70)	−.18	.11 (.10,.13)	.01
Noupoort (NC)	.69 (.65,.73)	−.17	.27 (.24,.29)	.10
Aberdeen (EC)	.62 (.60,.63)	−.14	.40 (.37,.44)	.30
Graaff-Reinet (EC)	.45 (.43,.47)	−.19	.31 (.27,.34)	.25
Victoria West (NC)	.81 (.80,.83)	−.11	.49 (.45,.52)	.29
Hopefield (WC)	.79 (.77,.81)	−.18	.69 (.66,.72)	.28
Belleville (WC)	.79 (.76,.83)	−.18	.69 (.66,.71)	.14
Calizdorp (WC)	.79 (.77,.80)	−.13	.73 (.68,.77)	.52
Uniondale	.64 (.63,.65)	−.10	.56 (.52,.59)	.45
Sutherland	.85 (.83,.87)	−.13	.72 (.67,.77)	.31

*Note.* NC = Northern Cape, EC = Eastern Cape, WC = Western Cape. Changes based on 100 simulations. Estimates of Coloured and non-Coloured support show 80% confidence intervals.



**Fig. 7** Fixed EI estimates of non-Coloured voting versus best guess estimates (South Africa data).

more or less correspond with common-sense predictions. In a comparison of Fig. 7 with Fig. 4, the improvement is quite remarkable.

In sum, including a covariate that measures the relative size of one of the population groups in the aggregate category can alleviate multimodality and aggregation bias caused by collapsing rows or columns. In the simulated data as well as in the example of Coloured voting in South Africa, the covariate greatly improved the reasonableness of estimates. This technique is easy to implement within the context of EI. It should thus be of value to researchers who suspect that the problems highlighted earlier in this article apply to their data but who are unable to run the types of models proposed by Rosen et al. (2001).

For some situations, however, this fix is not likely to work. For example, if the aggregate category is composed of three or more groups, it may be difficult to capture all multimodality and aggregation bias with a single covariate. While it is possible to include multiple covariates in EI, doing so places increasing strains on the information in the data, making estimation more difficult. In cases such as this, the approach in Rosen et al. (2001) is recommended.

## 7 Conclusion

The goal of this article has been to demonstrate that King's (1997) iterative approach to  $R \times C$  ecological inference can produce a set of problems unanticipated by *A Solution to the Ecological Inference Problem*. More specifically, under a set of general and not uncommon circumstances, a key step in the approach—combining rows or columns to produce residual or aggregate categories—can induce aggregation bias and/or multimodality and thus cause violations of model assumptions.

This increases the value of general  $R \times C$  approaches such as that proposed by Rosen et al. (2001). However, until these approaches become more practicable to the typical applied researcher, research using ecological inference techniques is likely to continue to rely on

the iterative approach proposed by King. Indeed, most current attempts to work with ecological data in political science use King's iterative approach. In recognition of this reality, I have proposed a quick fix that is easy to implement from within the EI framework. This fix appears to ameliorate the problem in the South African data that I have explored and also performs well in Monte Carlo analysis.

Ultimately, this article highlights once again the importance of paying close attention to the fit between the data and the assumptions of the model. This requires researchers to think critically and in new ways about the context of their data and the processes generating it. Standard EI diagnostics may do a poor job of picking up the problems I have outlined, and evaluating the assumptions of the model without regard to the new grouped data may miss critical violations. Finally, as demonstrated by the example of Coloured voting in South Africa, knowledge of what is reasonable, at both aggregate and local levels, can help reveal where and when the model goes wrong.

## References

- Chandra, Kanchan. 2004. *Why Ethnic Parties Succeed: Patronage and Ethnic Headcounts in India*. New York: Cambridge University Press.
- Ferree, Karen. 2002. *Voters and Parties in the Rainbow Nation: Race and Elections in the New South Africa*. Ph.D. thesis. Cambridge, MA: Harvard University.
- Independent Electoral Commission. 1994. *The South African Elections of April 1994*. Johannesburg: IEC.
- Johnson, R. W., and Lawrence Schlemmer. 1996. "National Issues and National Opinion." In *Launching Democracy in South Africa: The First Open Election, April 1994*, eds. R. W. Johnson and Lawrence Schlemmer. New Haven, CT: Yale University Press, pp. 74–107.
- King, Gary. 1997. *A Solution to the Ecological Inference Problem*. Princeton, NJ: Princeton University Press.
- King, Gary, Ori Rosen, and Martin Tanner. 1999. "Beta-Binomial Hierarchical Models for Ecological Inference." *Sociological Methods and Research* 28:61–90.
- King, Gary, Ori Rosen, Martin Tanner, and Alexander F. Wagner. 2003. "The Ordinary Election of Adolf Hitler: A Modern Voting Behavior Approach." Unpublished.
- Mattes, Robert. 1995. *The Election Book: Judgement and Choice in South Africa's 1994 Election*. Cape Town: Institute for Democracy in South Africa.
- Mattes, Robert, Hermann Gilliomee, and Wilmot James. 1996. "The Election in the Western Cape." In *Launching Democracy in South Africa: The First Open Election, April 1994*, eds. R. W. Johnson and Lawrence Schlemmer. New Haven, CT: Yale University Press, pp. 108–167.
- Posner, Daniel. Forthcoming. *Institutions and Identities: Regime Change and Ethnic Cleavages in Africa*. New York: Cambridge University Press.
- Reynolds, Andrew. 1994. "The Results." In *Election '94: The Campaigns, Results, and Future Prospects*, ed. Andrew Reynolds. New York: St. Martin's, pp. 182–220.
- Rosen, Ori, Wenxin Jiang, Gary King, and Martin A. Tanner. 2001. "Bayesian and Frequentist Inference for Ecological Inference: the  $R \times C$  Case." *Statistica Neerlandica* 55(2):134–156.
- Seekings, Jeremy. 1996. "From Independence to Identification." In *Now That We Are Free: Coloured Communities in a Democratic South Africa*, eds. Wilmot James, Daria Caliguire, and Kerry Cullinan. Cape Town: Institute for Democracy in South Africa, pp. 28–38.
- Wittenberg, Jason, and Jeffrey S. Kopstein. 2003. "Who Voted Communist? Reconsidering the Social Bases of Radicalism in Interwar Poland." *Slavic Review* 62:87–109.